09/077603

PCT/GB98 / 0 1 1 1 9

5

09/077603

Europäisches Patentamt | European Patent Office | Office européen des brevets

Bescheinigung | Certificate | Attestation

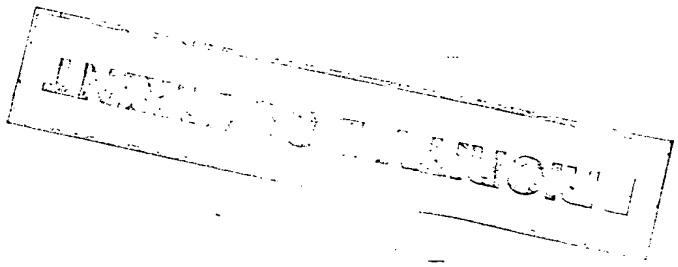
Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. | Patent application No. | Demande de brevet n°

97302616.4



Der Präsident des Europäischen Patentamts:
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

DEN HAAG, D
THE HAGUE, 27/04/98
LA HAYE

G. Monier

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.: 97302616.4
Demande n°:

Anmeldetag:
Date of filing: 16/04/97
Date de dépôt:

Anmelder:
Applicant(s):
Demandeur(s):
BRITISH TELECOMMUNICATIONS public limited company
London EC1A 7AJ
UNITED KINGDOM

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
Data summariser

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat:	Tag:	Aktenzeichen:
State:	Date:	File no.
Pays:	Date:	Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
G06F17/30

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du dépôt:

Bemerkungen:
Remarks:
Remarques:

DATA SUMMARISING SYSTEM

TECHNICAL FIELD

This invention lies in the field of systems, methods and apparatus for processing data information and more particularly in the field of systems, methods and apparatus for summarising data.

BACKGROUND TO THE INVENTION

Recent advances in technology have provided a vast increase in the availability of electronic data resources such as CD-ROM and in the volume of data available from systems such as the Internet. With this increase comes the associated problems of data overloading and the need for efficient data filtering systems.

Data summarisers are one element in this data filtering process. They operate to automatically produce a summary of data for instance so that a user of a system can determine if the data warrants further investigation.

According to a first aspect of the present invention there is provided a system for summarising data sets comprising:

a first data store for target data items;

means for dividing said data set into sections and for comparing each said section against said target data items;

means for calculating a ranking value for each said section dependent on the outcome of a said comparisons; and

means for compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

According to a further aspect of the present invention a method of summarising a data set input to processing apparatus having a data store for target information is provided; the method comprising the steps of

1) dividing said data set into sections;

2) comparing said sections against said target information;

3) calculating a ranking value for each said section dependent on the outcome of said comparison;

4) compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

According to a further aspect of the present invention there is provided apparatus for summarising a data set comprising:

a first data store for target data items;

means for dividing said data set into sections and for comparing each said section against said target data items;

means for calculating a ranking value for each said section dependent on the outcome of a said comparisons; and

means for compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

One advantage of the system, method and apparatus summary tools above is that the summaries are generated by reproducing sections of data within the data set that the tool believes is likely to be of interest to the user. Often such data is not the primary data detailed in a document and hence is often be missed by a prior-art summariser. This results in a data set being identified to a user as being relevant, and the user wading through the data set to locate the data items of specific interest.

Preferably the system or apparatus further comprises

means for identifying key data items within said data set and means for identifying a distribution pattern of said key data items within said data set;

calculating a second value for each said section dependent on the distribution of said key data items; and

means for modifying said ranking value dependent on said second value.

Preferably the method further comprises the steps of

5) identifying key data items within said data set;

6) identifying a distribution pattern of said key data items within said data set;

7) calculating a second value for each said section dependent on the distribution of said key data items;

8) modifying said ranking value dependent on said second value.

Refining ranking values according to distribution of key data items allows the summary that is generated to more accurately reflect the information content of the data being summarised. This is because the distribution of key data items can be used to estimate those data items reflecting the primary information content of the data being summarised.

Preferably said sections within the summary are ordered according to the order of their occurrence in the data set or according to their ranking value.

Preferably the second value is calculated for each section by determining a first score for each key data item in each section and summing said scores for each section; said first score calculated as the number of times the key data item of consideration occurs in the data less the number of times the key data item of consideration occurs in the section of consideration.

Preferably a second score is calculated for each key data item; said second score calculated by assigning a third value to each section of the data set; the third value corresponding to the position of the section within the data set and for each key data item performing the calculation of subtracting the third value of the first section in which said key data item occurs from the third value of the final section in which said key data item occurs; modifying said ranking values dependent on said second scores.

Preferably a third score is calculated for each key data item by identifying every pair of section in which key data items co-occur; for each pair of sections subtracting the lower second value from the higher second value and dividing the result by the second score; summing the third scores calculated for each section whereby calculating a fourth value for each section; modifying the rank value for each section dependent on the fourth value of each section.

BRIEF DESCRIPTION OF THE DRAWINGS

An information summariser according to an embodiment of the present invention will now be described, by way of example only, with reference to the accompanying figures, in which:

Figure 1 shows an information retrieval and processing system incorporating the information summariser;

Figure 2 shows a block diagram of the information summariser of Figure 1;

Figure 3 shows a flow chart of overall operation of the information summariser of Figure 1;

Figure 4 shows a flow chart of preferred steps for ranking sections of information according to their similarity to target information for use in the operation of Figure 3;

Figure 5 shows a flow chart of steps for calculating weightings for each section for use in the operation of Figure 3; and

Figure 6 shows a flow chart of steps for calculating the weightings attributable to each key word/phrase of information being summarised in an operation according to Figure 3.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to Figure 1, an information summarising tool according to an embodiment of the present invention may be built into a known form of information retrieval architecture, such as a client-server type architecture connected to the Internet.

In more detail, a customer, such as an international company, may have multiple users equipped with personal computers or workstations 405. These may be connected via a World Wide Web (WWW) viewer 400 in the customer's client context to the customer's WWW file server 410. An information summarising tool 120 may form an extension of the viewer 400, and may actually be resident on the WWW file server 410.

The customer's WWW file server 410 is connected to the Internet in known manner, for instance via the customer's own network 415 and a router 420. Service providers' file servers 425 can then be accessed via the Internet, again via routers.

Also resident on, or accessible by, the customer's file server 410 are an information retrieval tool 105 and two data stores, one holding target information and user profiles (the target information/ user profile store 430) and the other (the

intelligent page store 100) holding principally metainformation for a document collection.

The information retrieval tool 105 may be of a type known as a Jasper agent. Jasper agents are detailed in the applicants co-pending application PCT GB96/00132 which is incorporated herein by reference.

As described above, in the client-server architecture, the information retrieval tool 105 and the target information/user profile store both reside within the customer file server 410 where the summarising tool 120 may be resident

In one embodiment the summarising tool 120 may be built as an extension of a known viewer such as Netscape and operate to summarise WWW pages extracted by viewer 400. The summarising tool 120 generally receives as input, documents and/or pages located by either the search engines of the viewer 400 or by the information retrieval tool 105.

The summariser will typically divide the page/document into sections. These sections are typically sentences or paragraphs. These sections are then compared with the information stored in target information/user profile store 430 for a particular user.

On the basis of this comparison each section is assigned a value. This value is an indicator of the degree of similarity between the section and the target information/user profile information.

In one embodiment, a summary of the document/WWW page is generated by reproducing those sections having a value above/below a predefined threshold. These sections may be reproduced in the order in which they appeared in the document/page or they may be reproduced in value order.

The target information and profile information store detailed in file server 410 is preferably an intelligent store adapted to associate additional information with the key words, terms, phrases and other information specified by the user. An example of an intelligent target information/user profile information store is detailed in the applicants co-pending application PCT GB96/00132 referred to above.

Figure 2 is a block diagram of a preferred embodiment of the present invention. It comprises processing engine 10 which is adapted to receive text

information 15 as input. The preferred embodiment described has a further three inputs. These are target information/user profile information, stoplist information and stem information.

The target information\user profile information preferably contains details of information types that a user wishes to identify within the retrieved information. Such details include key words, key terms and key phrases and other information that server to summarise and/or define information of the type that the user wishes to identify and/or target.

The stop list information preferably contains lists of words such as typically commonly used words that typically do not make up keywords within a document. Such a list may be used to delete superfluous information from the document/page being summarised, thereby assisting to identify key words. The stop list may also contain information such as common phrases and terms. Alternatively, it may also contain specific key words, terms or phrases that a user requests be identified as key words.

Stem information contains a list of pre-fixes and suffixes which are used to reduce identified key words, terms or phrases to their basic form. For example assume that the word 'bounce' is a key word. The stem list preferably operates to reduce 'bouncing', 'bounced', 'bounces' and 'de-bounced' to the basic term 'bounce'.

Alternative embodiments of the present invention may use a natural language processing algorithm and/or system or some other technique known in the art to identify key words, terms or phrases within the information being summarised. According to one embodiment, this keyword analysis results in the production of a list that itemises the keywords appearing in each section.

Figure 3 is a flow chart identifying the steps detailed above for analysing the information and generating a summary.

According to Figure 4, one embodiment analyses the distribution of keywords throughout the document/page. This distribution is used to assign a weighting to each section. This weighting in turn is used to refine the order of sections having similar values assigned by the comparison process.

The distribution of keywords is used to measure the relevance of each section compared to the overall document/page content. It is assumed that those keywords with the widest distribution throughout the document/page are likely to be an accurate reflection of the document/pages information content. Other assumptions mat be utilised in this process. For example it may be assumed that keywords appearing at the beginning and end of a document form part of the introduction and conclusion and accordingly that they form an accurate reflection of the information content of the document/page. Similarly it may be assumed that the key words appearing at the beginning of a section are more significant than those at the end of a section.

Accordingly, sections containing information of greatest relevance to the target information/user profile information are assigned the highest ranking by the summarising tool. Sections of similar or identical ranking are ordered such that those sections with the highest weighting have their ranking increased over those sections with lower weightings.

According to one embodiment, detailed in figure 5, the list of key words generated by the keyword analysis is used as a representation of the keyword distribution. In this embodiment, each keyword is only accounted for once per section as it is assumed that section length is not an indication of the relative importance of the section to the information content of the document/page. Accounting for each keyword once per section helps to normalise the weightings against section length.

For each keyword in each section a value corresponding to the number of other sections in which the keyword appears is calculated. The weighting assigned to each section is equal to the sum of the values assigned to each keyword in the section.

Figure 6 details a value that is calculated for each keyword. This value is used by the process detailed in figure 7. In the preferred embodiment the calculation of figure 6 follows after the calculation detailed for figure 5.

The value calculated for each keyword in figure 6 is a further measure of keyword distribution within the page/document. The value is the number of sections separating the first and last occurrence of each keyword within the

page/document. It is assumed in using this value that keywords having wide separation within a document more accurately reflect the primary information content of a page/document than keywords with narrow separations.

The process of figure 7 is a modification to the weighting value calculated for each section as detailed in figure 5.

The process of figure 7 may be represented by the following section of pseudo-code:

for each sentence S

set x to zero

for each word

for every pair of sentences (i, j), where i>j, that the word occurs in

add $x = (i - j)/Ww$ to the weighting value of sentences S_i and S_j

where Ww is the value calculated for each key word in figure 6.

and where S_i and S_j are the weighting values calculated for each section in figure 5.

The following paragraphs will now be summarised by way of example and according to the operation of a preferred embodiment.

The cat sat on the mat.

A mat, a mat, my kingdom for a mat!

The dog also sat on the mat.

Both cat and dog sat on the mat.

The mat is on the floor.

The night was clear.

I counted the stars that night.

The dog sat on the floor.

The target information is the word "night".

Stoplisting, Stemming and Word Omission

Stoplisting and stemming is applied to the text and any words that only occur in one sentence are omitted. Repeat occurrences of a word in a sentence are ignored. This results in:

#	significant words in sentence
1:	cat sat mat
2:	mat
3:	dog sat mat
4:	cat dog sat mat
5:	mat floor
6:	night
7:	night
8:	dog sat floor

where the sentences are numbered in sequence. The position of words within sentences is insignificant.

Sentence Weighting

Each sentence, s, is given a weight, Ws , that is equal to the sum of the number of times each of its words occur in other sentences. For example, the last sentence has a weight of 6 ("dog" occurs twice elsewhere, "sat" three times, and "floor" once):

#	significant words in sentence	sentence weight, Ws
1:	cat sat mat	8
2:	mat	4
3:	dog sat mat	9
4:	cat dog sat mat	10
5:	mat floor	5
6:	night	1
7:	night	1
8:	dog sat floor	6

Sentence weights are used in step 5, below.

Word Weighting

Each word, w, is given a weight, W_w, that is equal to its greatest separation in terms of number of sentences. For example, the word "sat" has a weight of 7 (=8-1) since it occurs in sentences 1 and 8.

word, w	word weight, W_w
cat	3
sat	7
mat	4
dog	5
floor	3
night	1

Score Generation

A score is built up for each sentence as follows:

for each sentence S

set its score to zero

for each word

for every pair of sentences (i, j), where i>j, that the word occurs in

add (i - j)/Ww to the scores of sentences s_i and s_j

Consider for example sentence 8. The word "dog" has a weight of 5(=8-3). The occurrence of "dog" in sentences 3 and 4 contributes:

(8-3)/5 + (8-4)/5 (=1.8)

Repeating these operations for the words "sat" and "floor" yields the total score for the sentence as:

(8-3)/5	+	(8-4)/5	+	(8-1)/7	+	(8-3)/7	+	(8-4)/7	+	(8-5)/3	(=5.09 approx)
"dog"		"dog"		"sat"		"sat"		"sat"		"floor"	

It is preferred that links between words in widely-spaced sentences are favoured, as it is assumed that this may indicate that a concept contributes significantly to the content of a document.

Score Normalisation

The score for each sentence is normalised to the sentence's weight:

sentence score → sentence score / $\sqrt{W_s}$

Thus the score for sentence 8 becomes:

5.09/$\sqrt{6}$ (=2.08 approx).

It is preferable to perform this normalisation so that longer sentences do not get proportionatly higher scores than shorter sentences. It has been found that dividing by $\sqrt{W_s}$ provides preferred results to dividing by W_s alone.

Score Skewing

The score for each sentence is adjusted according to the position of the sentence within its paragraph, with earlier sentences being favoured more highly.

The scheme used is:

sentence 1: score multiplied by 1.2

sentence 2: score multiplied by 1.1

sentence 3: score multiplied by 1.05

sentence 4: score multiplied by 1.025

etc

Thus the score for sentence 8 becomes:

2.08*1.05 (=2.18 approx)

because sentence 8 is the third sentence in its paragraph.

This skewing operates on the assumption that the most significant information in a paragraph is often found near its start.

A similar skewing is also applied to each paragraph:

paragraph 1:score multiplied by 1.2

paragraph 2:score multiplied by 1.1

paragraph 3:score multiplied by 1.05

paragraph 4:score multiplied by 1.025

etc

Thus the score for sentence 8 becomes:

2.18*1.1 (=2.39 approx)

because sentence 8 is in the second paragraph.

Sentence rating (fine gradation)

Applying the previous steps, yields the following (approximate sentence scores:

#	sentence score
1:	2.65
2:	1.16
3:	1.61
4:	1.90(3)
5:	1.90(2)
6:	1.32
7:	1.21
8:	2.39

Rating the sentences according to score gives:

#	sentence rating (fine gradation)
1:	8
2:	1

3: 4

4: 6

5: 5

6: 3

7: 2

8: 7

In this example a separate rating is assigned to each sentence: sentence 1 is the most significant (rating 8), and sentence 2 the least significant (rating 1). These ratings may be used as they stand to provide summaries of all possible lengths, by varying a threshold rating, and only including sentences with ratings at or above the threshold.

In some embodiments the gradation in this rating may be too fine in which case a coarser rating system, as described below, may be used.

Sentence Rating (coarse gradation)

With a coarse gradation rating scheme, the number of unique sentences ratings is collapsed into a smaller number, so that summaries of approximately 1/2, 1/4, 1/8 etc of the original document length (with a lower limit of two sentences) are produced.

For the example text, the mapping from coarse rating to fine rating is:

coarse: 8 7 6 5 4 3 2 1

fine: 3 3 3 2 2 1 1 1

giving the coarse sentence ratings as:

sentence rating (coarse gradation)

1: 3

2: 1

3: 2

4: 3

5: 2

6: 1

7: 1

8: 3

Thus selecting a threshold rating of 2 would produce a summary containing sentences 1, 3, 4, 5 and 8.

Profile Handling

The ratings of all sentences containing words or phrases that match the user profile are increased sufficiently to exceed the scores of all other sentences. In the case where there is more than one word or phrase in the user profile, all sentences containing N + 1 matches to the user profile have their scores increased sufficiently to exceed the scores of all sentences containing N matches to the user profile. Multiple occurrences of a profile's word or phrase within a sentence are ignored.

For example, with a user profile of "night, star", the rating of sentence 6 (containing "night") is increased from 1 to 4, and the rating of sentence 7 (containing both "night" and "star") is increased from 1 to 5. Differences in original gradation are preserved when promoting ratings to take account of the user profile.

Now the coarse sentence ratings as:

#	sentence rating (coarse gradation)
1:	3
2:	1
3:	2
4:	3
5:	2
6:	4
7:	5
8:	3

Thus selecting a threshold rating of 3 would produce a summary containing sentences 1, 4, 6, 7 and 8.

CLAIMS

1. A system for summarising data sets comprising:

a first data store for target data items;

means for dividing said data set into sections and for comparing each said section against said target data items;

means for calculating a ranking value for each said section dependent on the outcome of a said comparisons; and

means for compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

2. A system as claimed in claim 1 further comprising:

means for identifying key data items within said data set and means for identifying a distribution pattern of said key data items within said data set;

calculating a second value for each said section dependent on the distribution of said key data items;

means for modifying said ranking value dependent on said second value.

3. A system as claimed in claims 1 or 2 wherein said sections within the summary are ordered according to the order of their occurrence in the data set.

4. A system as claimed in claims 1 or 2 wherein said section within said summary are ordered according to their ranking value.

5. A system as claimed in any one of claims 2 to 3 wherein said second value is calculated for each section by determining a first score for each key data item in each section and summing said scores for each section; said first score calculated as the number of times the key data item of consideration occurs in the data less the number of times the key data item of consideration occurs in the section of consideration.

6. A system as claimed in any one of claims 2 to 5 wherein a second score is calculated for each key data item; said second score calculated by assigning a third value to each section of the data set; the third value corresponding to the position of the section within the data set and for each key data item performing the calculation of subtracting the third value of the first section in which said key data item occurs from the third value of the final section in which said key data item occurs; modifying said ranking values dependent on said second scores.

7. A system as claimed in claim 6 wherein a third score is calculated for each key data item by identifying every pair of section in which key data items co-occur; for each pair of sections subtracting the lower second value from the higher second value and dividing the result by the second score; summing the third scores calculated for each section whereby calculating a fourth value for each section; modifying the rank value for each section dependent on the fourth value of each section.

ABSTRACT

DATA SUMMARISING SYSTEM

According to a first aspect of the present invention there is provided a system for summarising data sets comprising:

a first data store for target data items;

means for dividing said data set into sections and for comparing each said section against said target data items;

means for calculating a ranking value for each said section dependent on the outcome of a said comparisons; and

means for compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

According to a further aspect of the present invention a method of summarising a data set input to processing apparatus having a data store for target information is provided; the method comprising the steps of

1) dividing said data set into sections;

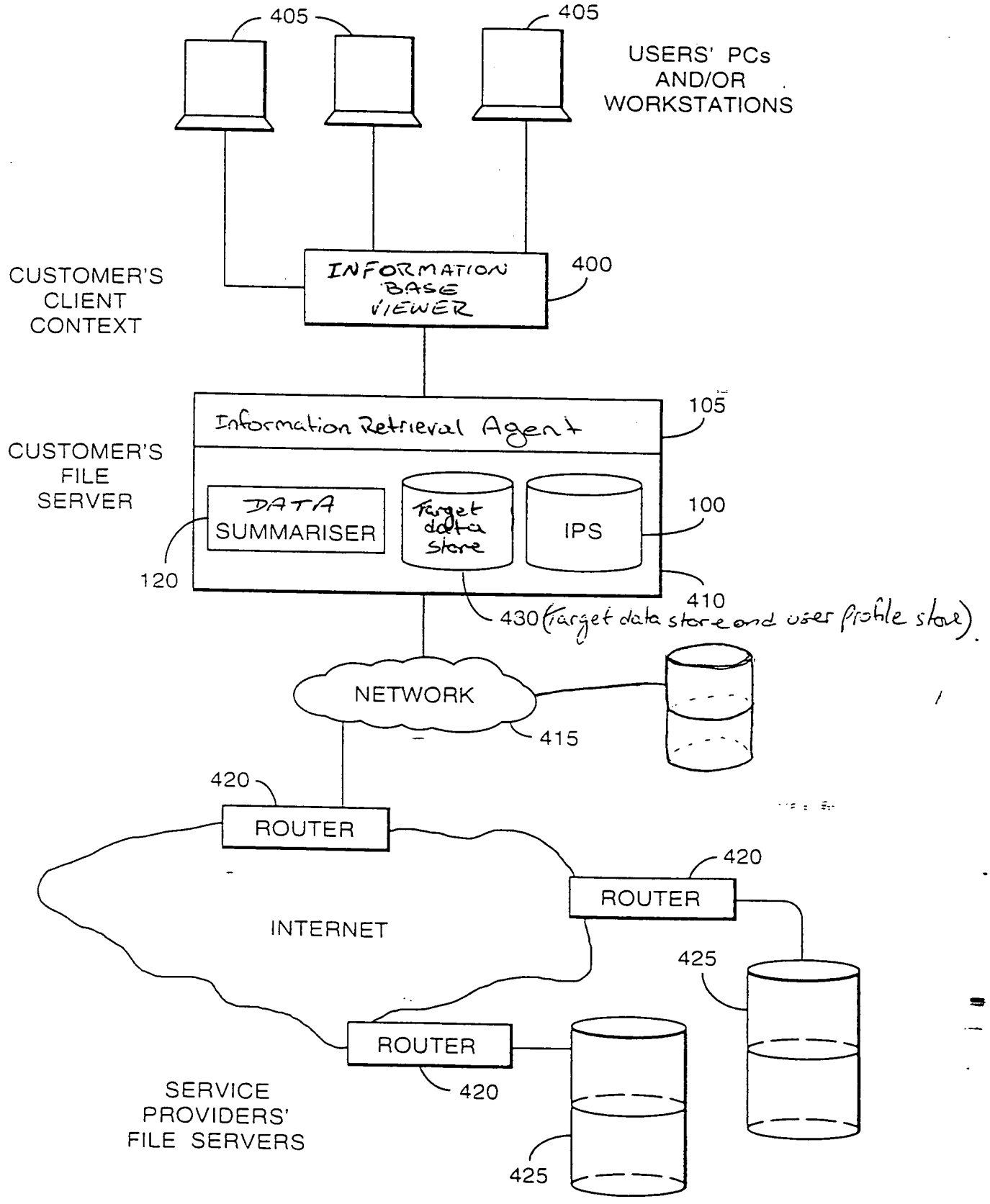
2) comparing said sections against said target information;

3) calculating a ranking value for each said section dependent on the outcome of said comparison;

4) compiling a summary of the data set from sections having a ranking value past a pre-determined threshold value.

Figure (2)

Fig.1.


405
405
USERS' PCs AND/OR WORKSTATIONS
CUSTOMER'S CLIENT CONTEXT
INFORMATION BASE VIEWER
400
CUSTOMER'S FILE SERVER
Information Retrieval Agent
105
DATA SUMMARISER
Target data store
IPS
100
120
410
430 (Target data store and user profile store)
NETWORK
415
420
ROUTER
INTERNET
420
ROUTER
425
ROUTER
420
SERVICE PROVIDERS' FILE SERVERS
425

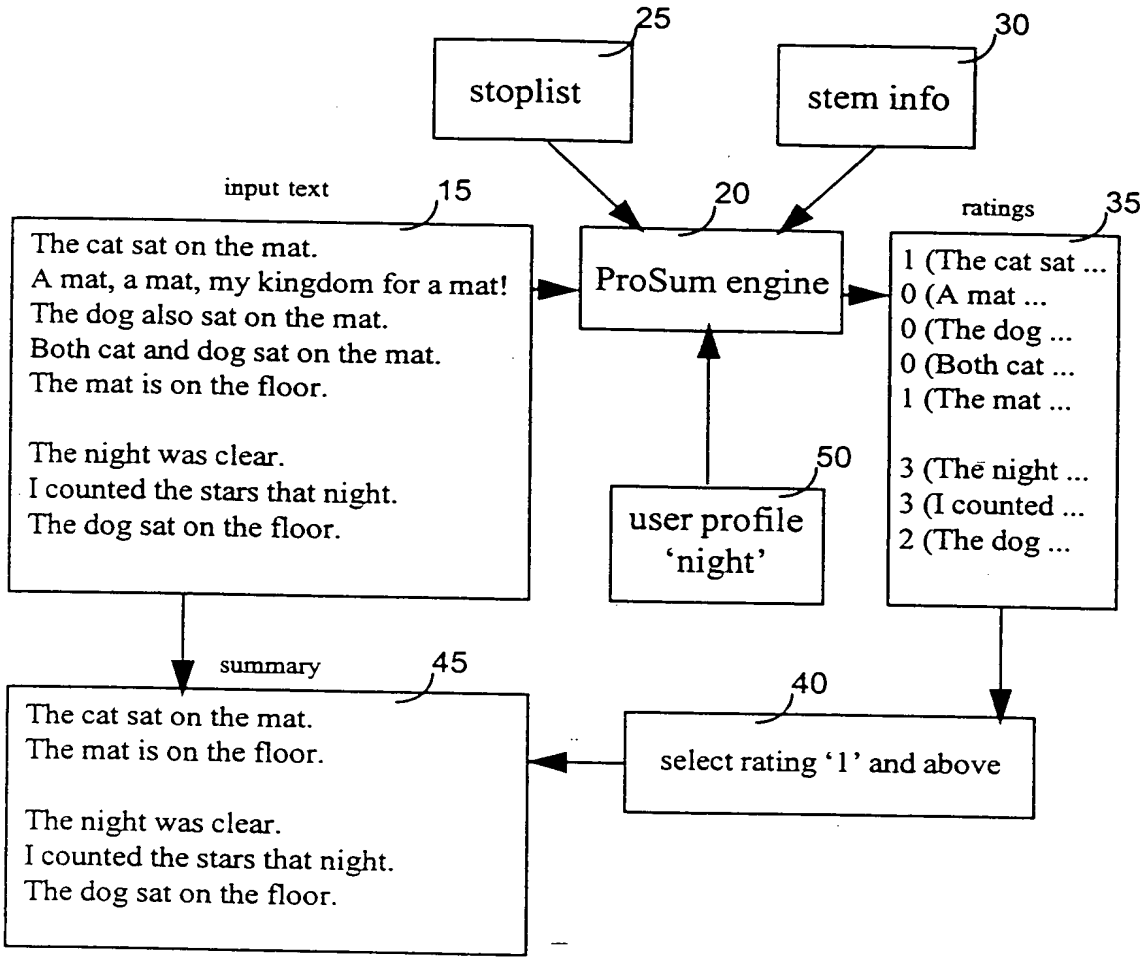

25
stoplist
30
stem info
input text
15
20
ratings
35
The cat sat on the mat.
A mat, a mat, my kingdom for a mat!
The dog also sat on the mat.
Both cat and dog sat on the mat.
The mat is on the floor.
The night was clear.
I counted the stars that night.
The dog sat on the floor.
ProSum engine
1 (The cat sat ...
0 (A mat ...
0 (The dog ...
0 (Both cat ...
1 (The mat ...
3 (The night ...
3 (I counted ...
2 (The dog ...
50
user profile
'night'
summary
45
40
The cat sat on the mat.
The mat is on the floor.
The night was clear.
I counted the stars that night.
The dog sat on the floor.
select rating '1' and above

Figure 2

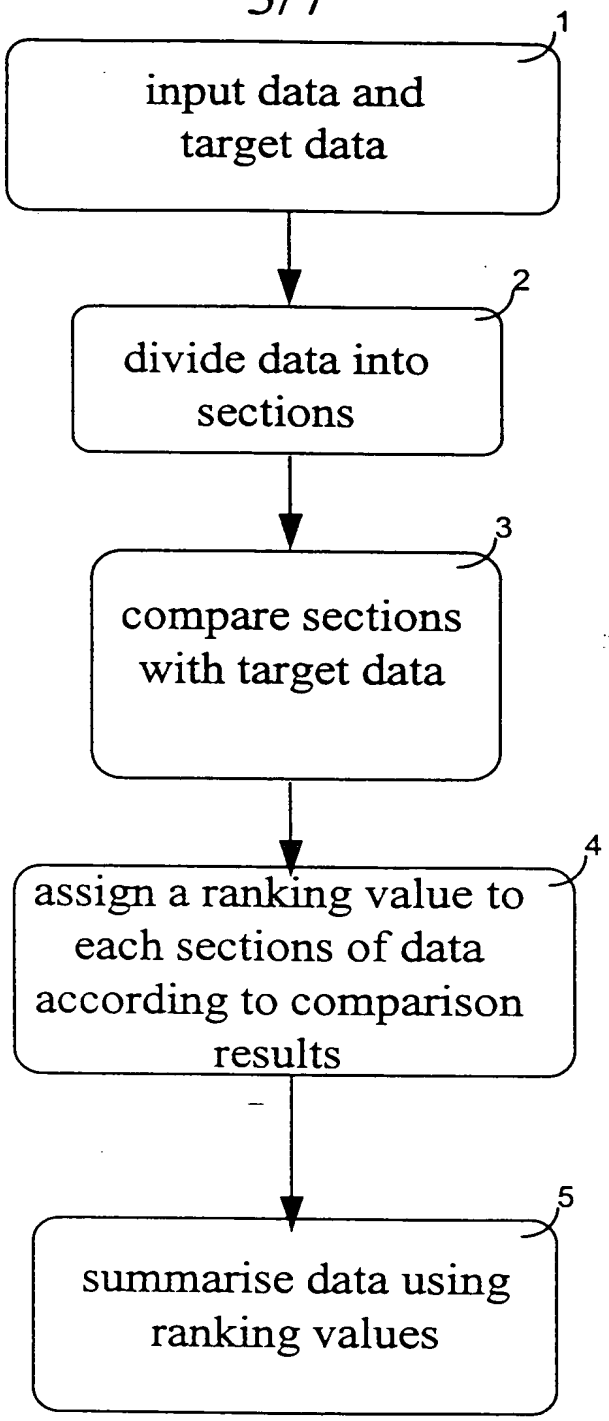

1
input data and target data
2
divide data into sections
3
compare sections with target data
4
assign a ranking value to each sections of data according to comparison results
5
summarise data using ranking values

Figure 3

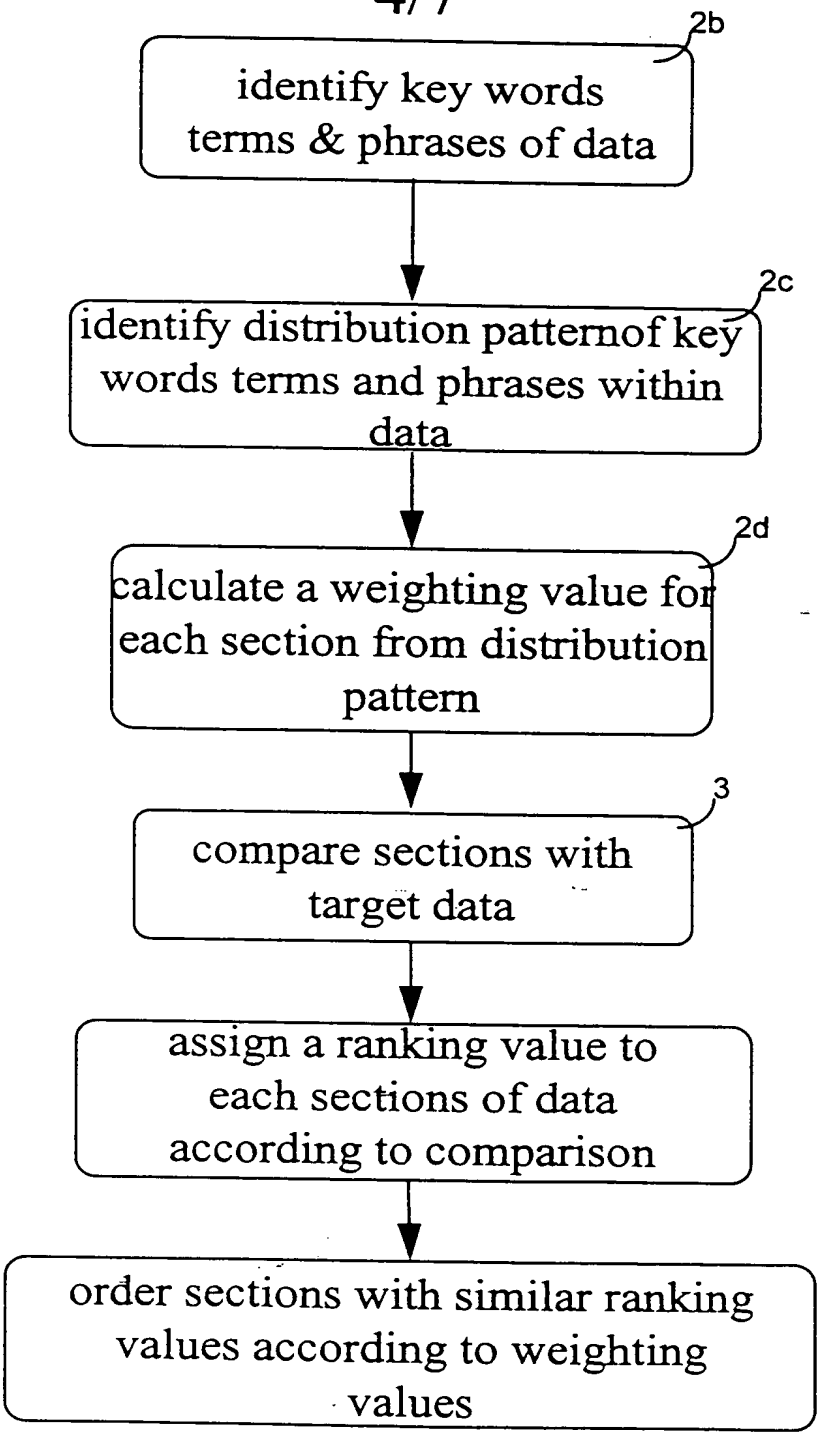

2b
identify key words terms & phrases of data
2c
identify distribution patternof key words terms and phrases within data
2d
calculate a weighting value for each section from distribution pattern
3
compare sections with target data
assign a ranking value to each sections of data according to comparison
order sections with similar ranking values according to weighting values

Figure 4

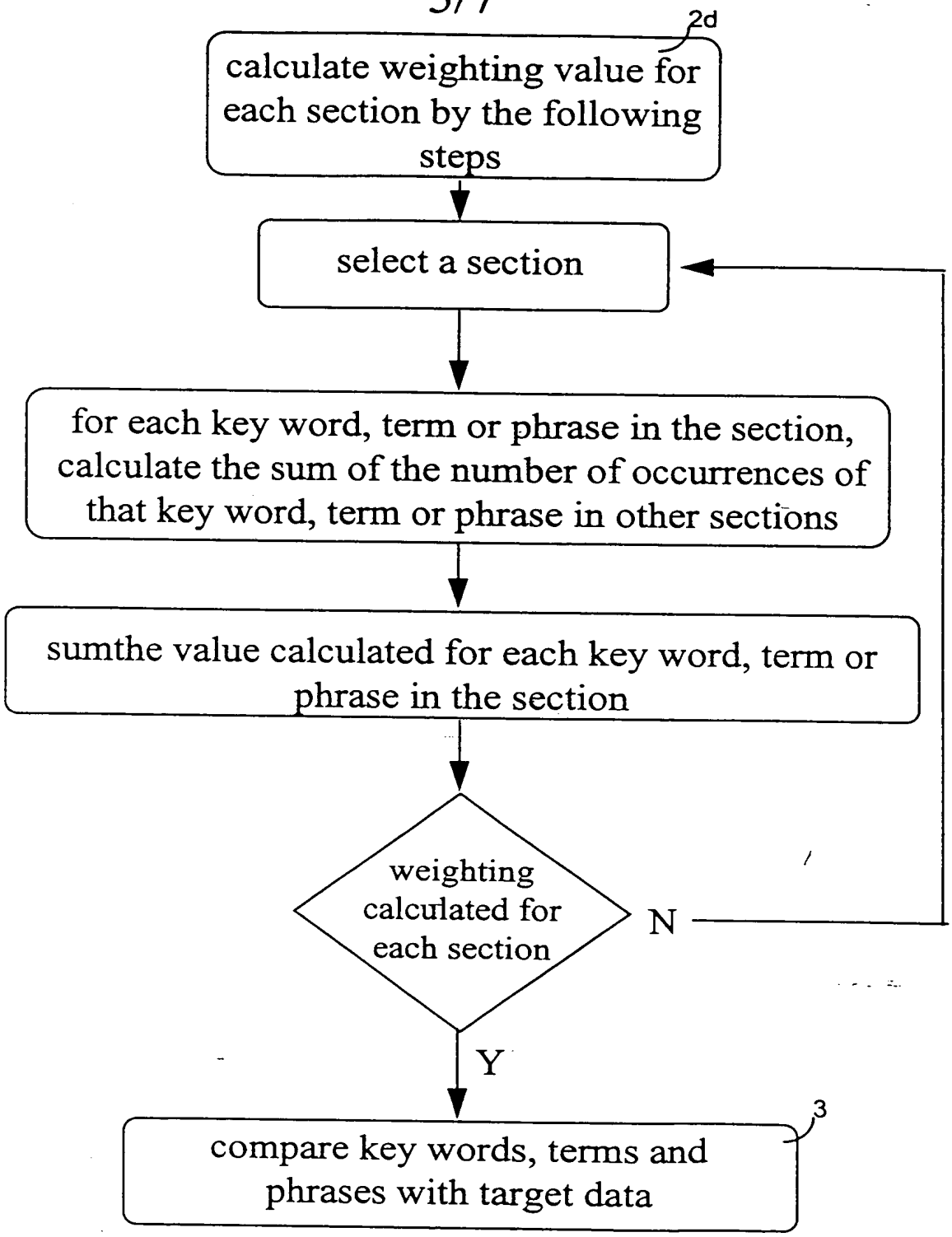

2d
calculate weighting value for each section by the following steps
select a section
for each key word, term or phrase in the section, calculate the sum of the number of occurrences of that key word, term or phrase in other sections
sumthe value calculated for each key word, term or phrase in the section
weighting calculated for each section
N
Y
3
compare key words, terms and phrases with target data

Figure 5

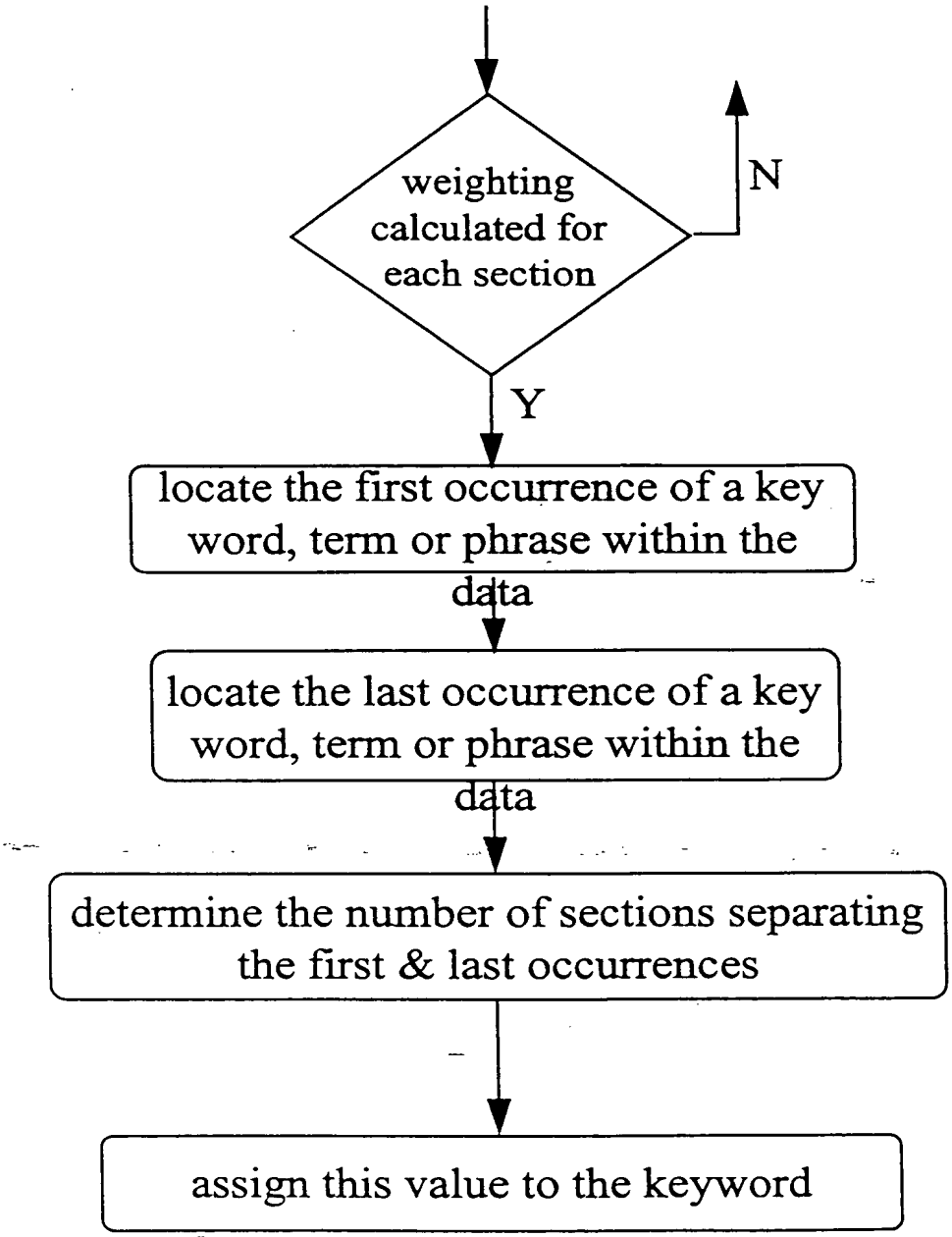

weighting calculated for each section
N
Y
locate the first occurrence of a key word, term or phrase within the data
locate the last occurrence of a key word, term or phrase within the data
determine the number of sections separating the first & last occurrences
assign this value to the keyword

Figure 6

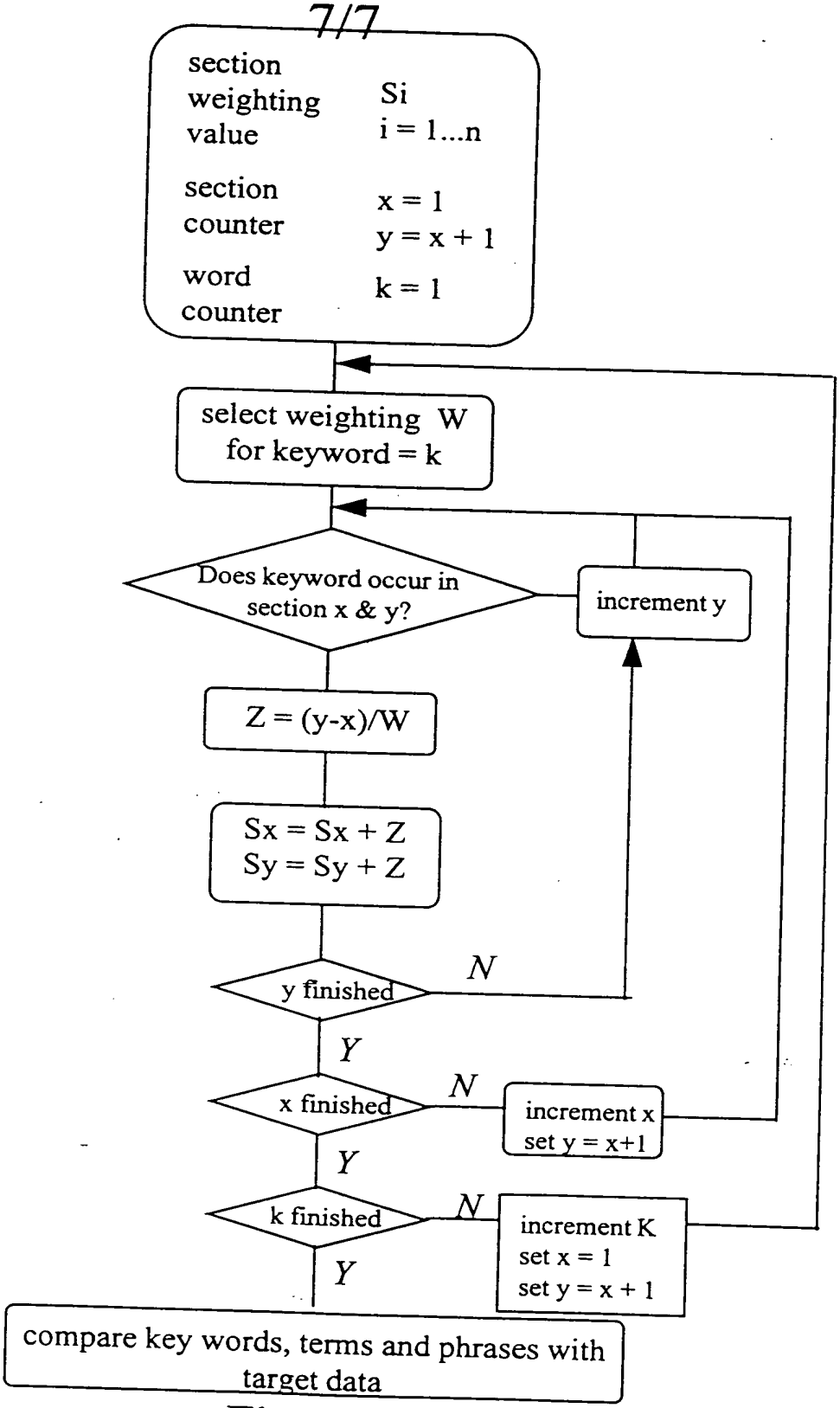

section weighting value
Si
i = 1...n
section counter
x = 1
y = x + 1
word counter
k = 1
select weighting W for keyword = k
Does keyword occur in section x & y?
increment y
Z = (y-x)/W
Sx = Sx + Z
Sy = Sy + Z
y finished
N
Y
x finished
N
increment x
set y = x+1
Y
k finished
N
increment K
set x = 1
set y = x + 1
Y
compare key words, terms and phrases with target data

Figure 7